Hybrid HMM/SVMを利用した筋電位に基づく日本語黙声認識

Hybrid HMM/SVMを利用した
筋電位に基づく日本語黙声認識

筑波大学
図書館情報メディア研究科

2013年　3月

大内　慶久

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究背景

近年，音声認識技術の発展により，音声入力インタフェースを用いたサービスが身近なものとなりつつある [1][2]. 音声入力インタフェースでは人間が発声する自然言語をコンピュータが認識するために，慣れが必要なキーボード入力に比べて自然で直感的といえ，入力に要する時間も少ない．しかし，図書館など公共の場所で静粛性が求められる場合において声を出して発話することが周囲の人々の迷惑となったり，第三者に声が伝わることで秘匿性に難があったりして，その利用に際して問題となることがある．また，音声発話という機能が失われた場合には，代替システムが必要となる．

そこでこれらの問題を解決するために，音声を用いることなく発話内容を認識する技術が研究されている．その技術の一つとして，画像認識を利用して発話時の口唇周辺の動作を読み取る読唇がある [3][4]. 読唇は，話者に対して非接触で行うことが可能であるが，画像取得のためにカメラの設置が必要であるといった制約がある．また，口唇位置の検出や環境変化に頑強な特徴抽出などに課題があり，これまでに単語認識に関してはその有効性が示されているものの，母音や子音の音素といったより小さな音韻での認識は考慮されていない．

これとは別の手法として，筋収縮に伴って生じる筋電位信号を利用した筋電インタフェースに関する研究がある．筋電インタフェースでは，皮膚表面に装着した電極を用いて筋電位を観測し，発話に伴う筋の動きに対して特徴的なパターンを得ることで，その内容を認識する．日本語に関しては，吉川らが，表情筋及び頸部で計測した筋電位を基に母音認識を行い，k-近傍法，ベイズ決定則，ニューラルネットワーク，サポートベクトルマシン（Support Vector Machine:SVM），隠れマルコフモデル（Hidden Markov Model:HMM）の五つの認識手法による認識精度を比較している [9]. また福田らは，ニューラルネットワークとHMM を組み合わせて二十単語での単語認識を行っている [7]. 真鍋らは広くユーザに受け入れられる測定方法として，指輪型電極による筋電位計測に基づいた母音認識を試みている [6]. 日本語以外では，英語 [10] やアラビア語 [11] の認識に関する研究も行われている．

さて，音声を用いずに発話内容の認識を行う場合も，大規模語彙を想定する場合には音声認識と同様，音素単位の認識を考慮する必要がある．日本語では子音が単独で存在せず，基本となる五母音と子音を組み合わた五十音（濁音，半濁音も含む）を音韻として最小の単位とみることができる．つまり，母音の音素の前に子音の音素が存在するため，筋電位を見た場合には，子音部分を示す筋電位は母音部分の前に存在すると考えられる．筋電位を用いた子音の認識に関しては，日本語五十音で一音ごとに計測された筋電位を観察することで，その認識可能性に関する調査を行った研究はあるが，明確な結論には至っていない [12][13]. しかしこの調査では，子音から母音に移る際の変化に対応する波形がしばしば観察されていることから，このような短い時間変化を表現できる認識モデルを設計する

ことで，子音を含む音韻の認識が可能であると考えられる．

吉川らの先行研究 [9] の五種類の認識器ごとの比較をみると，日本語五母音では HMM, SVM のときにそれぞれ認識率が 90% を超える高い認識率が得られている．HMM は通常の音声認識で用いられ，時系列信号のモデル化に有効な手法として知られるものである．SVM は，文字認識や画像認識などの応用分野において，高い認識性能を示すことが確認されているが [25]，筋電位のような時系列でパターンの長さが伸縮する信号に対しては，何らかの対処が必要である．先行研究 [9] では入力区間内で投票処理を行って，最も多く認識結果として出力されたものをその区間全体の結果としているが，この方法では時系列パターンの時間変化を考慮することはできない．そこで，時系列信号を複数の定常信号源の連鎖によってモデル化することにより高い認識性能を示す HMM の枠組みを応用し，複数の SVM を接続して構成する連鎖モデルが提案されている．しかし，連鎖モデルにおける遷移確率がすべて等しくなければならないといった制限があるために時系列パターンの時間伸縮を十分に表現できているとはいえない [17]．しかし，このように認識性能の高い SVM において時系列のパターンをうまく表現する認識モデルを設計することができれば，筋電インタフェースにおいても従来の手法より高い認識性能が期待でき，先に述べた子音の認識にも有効であると考えられる．

以下では，発声しない発話動作，いわゆる口パクに関して，黙声と表現することとする．

1.2　研究の目的と提案

本研究の目的は，日本語黙声認識を実現するために，日本語五十音の認識を行うことである．そのために，従来の認識手法を改良した認識モデルを日本語五十音の黙声認識に適用することを提案する．そして，評価実験によってその有効性を明らかにするとともに，子音に関する認識の可能性を考察する．具体的には，SVM を時系列パターンに適応するために，時間的な状態遷移を表現する連鎖モデルを利用するが，学習サンプルと連鎖モデルの各状態との対応関係を決定するために，HMM を利用した Viterbi アラインメントを用いる．Viterbi アラインメントでは，各状態に分割された区間のデータは定常的であることが期待でき，この各状態ラベルをクラスラベルとして SVM の学習を行う．そして，各状態間を遷移する頻度を遷移確率とし，これによって各状態及び状態を連結させたパターンの時間伸縮にも対応することが可能となる．SVM を連鎖させる際には先行研究 [17] と同様に，各状態で SVM の事後確率を利用した確率モデルとして構築するため，HMM の各状態での出力確率を SVM による事後確率に置き換えたような構成として記述でき，認識の際も HMM と同様に Viterbi アルゴリズムを利用する．本稿では，この認識モデルを Hybrid HMM/SVM と呼ぶこととする．

1.3　本論文の構成

本稿は全 6 章から構成される．第 1 章では序論として研究の背景と本研究の目的を示した．第 2 章では本研究に関連する基本原理・基礎技術として筋電位及び発声器官，そして従来の音声認識技術に関して述べる．第 3 章では筋電インタフェースシステムに関して主に信号計測部と特徴抽出部に焦点を当ててその概要を示す．第 4 章では本研究のシステム

において学習・認識部に用いる手法，HMM と SVM，そして Hybrid HMM/SVM について述べる．第 5 章では第 3 章及び第 4 章で述べたシステムを用いた日本語五十音の黙声認識実験を通してモデルの評価を行う．そして最後に，第 6 章で本研究全体についての結論と今後の展望を示す.

第2章　基本原理と基礎技術

2.1　筋電位信号

本節では，筋電位の発生メカニズムについて説明した後，観測される筋電位の性質について述べる．

2.1.1　生体電気の概要

人間が思考したり，行動したりするその過程において，電気信号が使われていることが知られている．人間の中枢神経系は脳と脊髄から成り，これらを構成する神経細胞（ニューロン）から発生した電気信号が活動電位と呼ばれるデジタル信号として，次のニューロンへと伝達される．ニューロンには，活動電位を次のニューロンに送る軸索という長い突起と，樹状突起という軸索入力を受ける二種類の突起がある．この二種類の突起はシナプスという微細構造で接合される．軸索を伝わってきた活動電位は，シナプスにおいて神経伝達物質と呼ばれる化学物質の放出を誘導し，これが樹状突起の受容体に結合することでシナプス電位というアナログ信号を次のニューロンに引き起こす．たくさんのシナプス入力を受けてシナプス電位が重なり，それがある閾値に達するとニューロンが興奮して，活動電位というデジタル信号がニューロンに発生する．そしてこの信号がまた軸索を通り，次々と伝達される．

2.1.2　筋電位の発生

筋活動を引き起こす基本単位を運動単位といい，一つの運動単位は一つのα運動ニューロンとそれが支配する数百から数千の筋線維の集まりから成る．そして，一つの筋肉は多数の運動単位で構成される．α運動ニューロンは中枢から運動指令を受け，α運動ニューロンで発生した活動電位が軸索末端に到達すると，神経伝達物質を放出する．筋線維の細胞はニューロンと同様の性質を持っており，神経伝達物質を受け取って興奮すると，活動電位が発生する．これによって，筋線維が収縮してその両端に張力が発生する．ここで筋線維で発生する活動電位を筋電位と呼ぶ．運動単位は興奮するかしないかのデジタルな振るまいをするので，筋張力は興奮する運動単位の数と興奮頻度によって変化し，次のように決定される [22].

$$\text{筋張力} = \sum(\text{運動単位} \times \text{興奮頻度}) \tag{2.1}$$

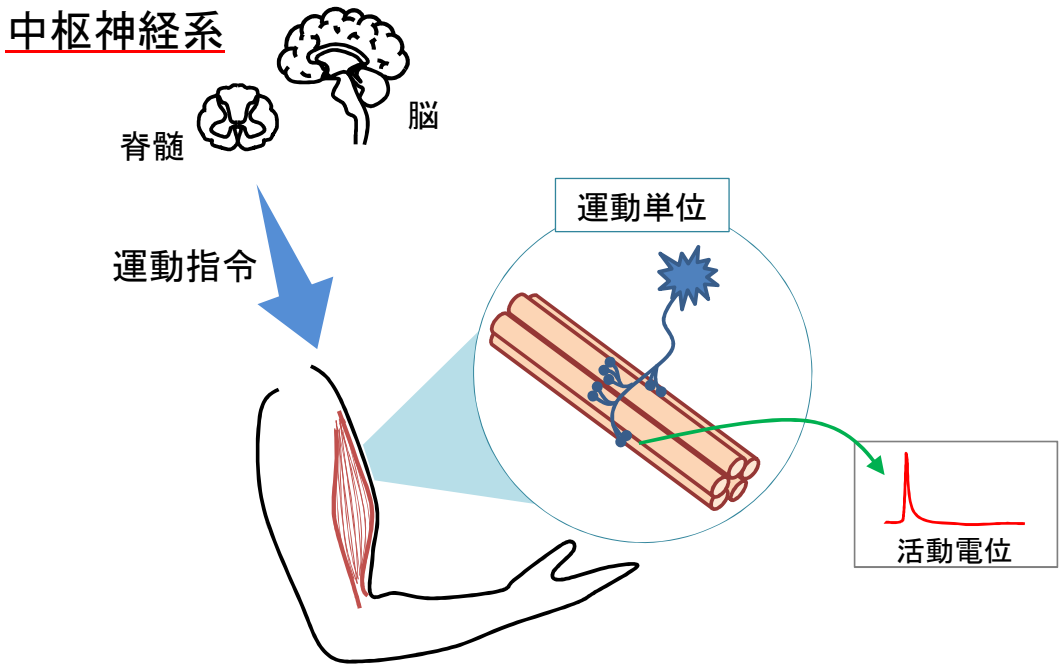

図 2.1: 筋電位の発生

2.1.3　筋電位の計測

筋電位の計測には針電極を用いるものと，表面電極を用いるものがある．針電極は皮膚から筋肉に刺入して用い，針筋電位信号として個々の運動単位の活動を計測することができるが，人体に対して侵襲性が高いために対象者の身体的な負担が大きく，主に神経筋疾患の診断などの臨床分野で主に用いられる．表面電極は皮膚表面から計測を行うため，表面筋電位信号は複数の運動単位で発生した筋電位が時間的・空間的に重畳した干渉波として計測される．また，筋電位により筋収縮が誘発されるために，筋電位は筋収縮よりも 30-100[ms] 先立って観測される．つまり，筋電インタフェースでは動作に先行する筋電位信号を利用するために，時間的に遅れのない動作認識が実現できる可能性がある．

2.2　音声生成と音声器官

音声生成のプロセスは，音源，調音，放射の三段階から成る．このうち，発声の際に言語音である母音及び子音を生成することを調音といい，声道の形を変化させて気流に影響を与えることによって行う．人間の発声器官の構造は，肺，気管，喉頭，鼻腔，口腔などから成っており，これらは全体として一つの連続した管を成している．ここで，口腔より上の部分は声道と呼ばれ，顎，舌，口唇などの調音器官を動かすことにより様々な形に変化する．子音の調音では，主に唇，舌尖，舌端，後舌面が用いられる．母音の場合には舌の最も高い点の位置と唇の構えを変化させて発声が行われる．図 2.3 は英語子音について，調音の位置と調音の型から分類したものである．また，図 2.4 は日本語母音に関して，舌の前後位置及び音響的な高さの観点から体系化したものを示した図である．

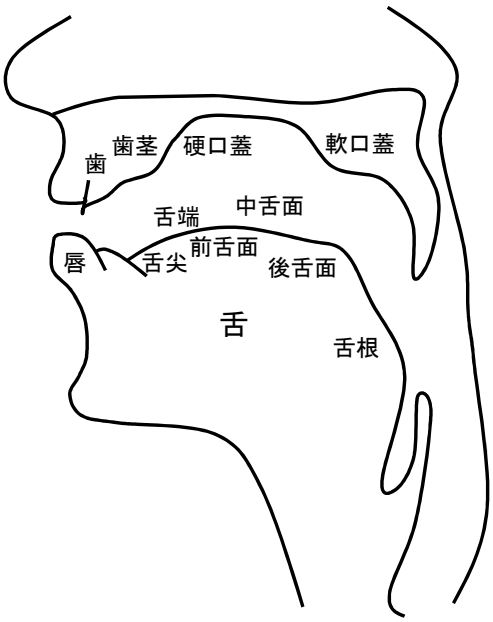

図 2.2: 声道の主要部分

調音の型 ＼ 調音の位置	両唇音	唇歯音	歯音	歯茎音	後部歯茎音	硬口蓋音	軟口蓋音
	唇音		舌頂音			後舌面音	
鼻音	m			n			ŋ
破裂音	p　b			t　d			k　g
摩擦音		f　v	θ　ð	s　z	ʃ　ʒ		
接近音				ɹ		j	w
側面接近音				l			

図 2.3: 調音の位置と型による英語子音の分類

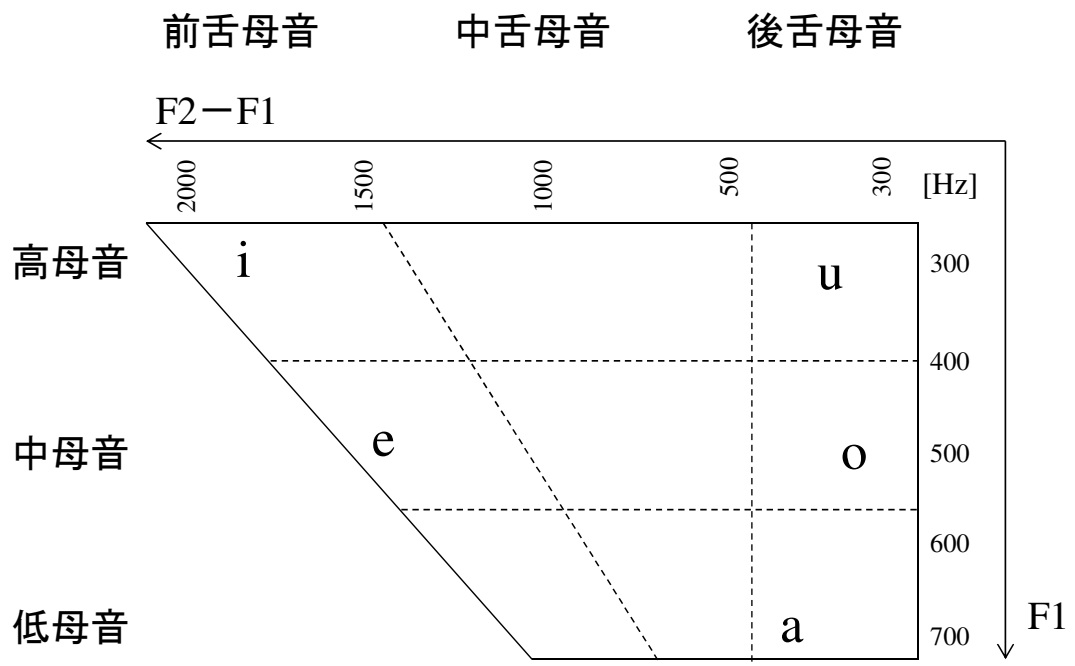

図 2.4: 日本語の母音の性質を示す母音図

2.3　音声認識技術

音声認識という言葉は，我々が他人の話す言葉を聞き取るというような意味で使われるが，工学の分野ではコンピュータによる音声の自動認識を指す．そして現在の音声認識システムを大別すると，一音ずつ区切って入力する音韻認識，一単語ずつ区切って入力する単語認識，句や文単位の認識，そして区切りなく発話された複数の文を入力とする連続音声認識がある．図 2.5 に示すようなシステム構成で実現される．

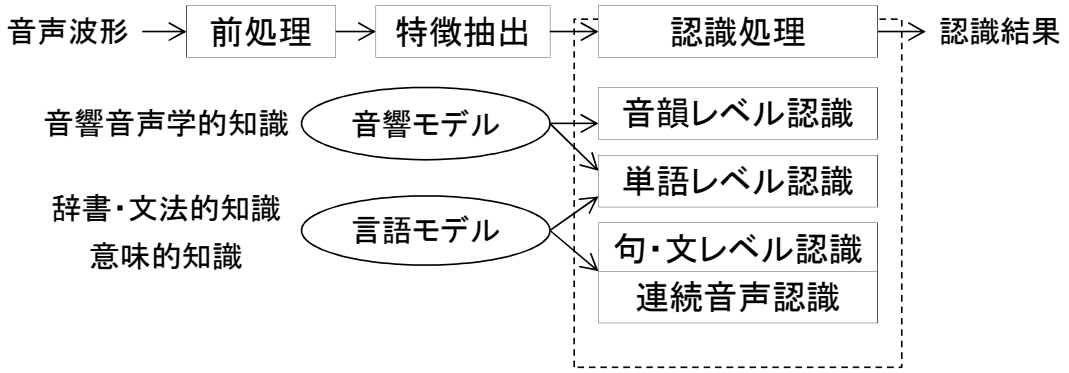

図 2.5: 連続音声認識システムの構成

2.3.1　特徴抽出

連続音声認識システムでは，時系列で与えられる音声信号のどこからどこまでがどの音に対応するかが分からないので，音声を分析することによって短時間で区切った特徴ベクトル系列を抽出し，これを認識に利用する．第一次的な特徴量としては次のようなものが挙げられる．

- 周波数パワースペクトル包絡
- 声道断面積関数
- ホルマント
- 音声波形の振幅
- 基本周波数
- 無音区間長
- 音韻区間長・発声速度

最近の認識システムでは，上記の特徴量のうちパワースペクトルに注目し，韻律に関する特徴量を補助的に利用するものが多い．パワースペクトル包絡に関連した変量は，ケプストラム，メルケプストラムとして表現されて多くのシステムで利用され，大語彙連続音声認識システム Julius[28] でもメルケプストラムが組み込まれている．

2.3.2　音韻認識と単語認識への拡張

特徴抽出を行った後は音韻認識が行われる．音韻認識や単語認識などは基本的に同じ形式で扱うことができ，対象を音素ないし単音とするか，単語とするかという点のみが異なる．また，単語認識や文単位の認識は，それよりも小さい単位で認識を行った結果を利用する方法で拡張していくことができる．その場合には音素や単音もしくは単語を一つのクラスに対応付け，単語は音韻の系列，文は単語の系列として連結したネットワーク構造で表現される．

音韻認識は以下のような手順で行われる．

1. 音韻の表現方法の決定（音素・単音など）
2. 音声サンプルの集合からの音韻の標準パターンの推定
3. 未知入力サンプルに対して，各標準パターンからの距離の算出

これらを統計的方法を用いて確率モデルで表現したものを音響モデルと呼び，音韻認識では音韻 ω を発声したときに，特徴ベクトル系列 $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$ がどれくらいの確率で観測されるかを表した条件付き確率 $P(\boldsymbol{O}|\omega)$ を与える．認識においては，特徴ベクトル系列 $\boldsymbol{O}$ を観測したときに，事後確率が最大となる音韻 ω を出力する．近年実用化されている音声認識システムの音響モデルでは，HMM が用いられることが多い．HMM については第 4 章で詳しく説明する．

ここでは，さらに音韻認識を単語認識へと拡張することを考える．単語は音韻系列$\omega_1, \omega_2, ..., \omega_n$ で表現されるため，各音韻の音響モデルの組み合わせ系列の中から，特徴ベクトル系列$\boldsymbol{O}$に対して事後確率が最大となる音韻系列を算出することにより，単語認識が可能となる．文単位の認識へも同様に，文を単語系列として表現して拡張していけばよい．またこのとき，単語が日常見かけるかどうか，もしくは単語系列が文法的に誤っているかどうかなどを言語モデル $P(\omega_1, \omega_2, ..., \omega_n)$ として与えることもできる．

第3章　筋電インターフェース

筋電インタフェースは，身体機能の拡張･増幅を行うロボットスーツ [16] や筋電義手 [20] の開発に利用されている．ここでは，筋電インタフェースでの処理について述べるとともに，本研究において参考にしたシステム [9][20] における特徴抽出手法について述べる．

3.1　処理の流れ

図 3.1 に，筋電インタフェースの処理の流れを示す．

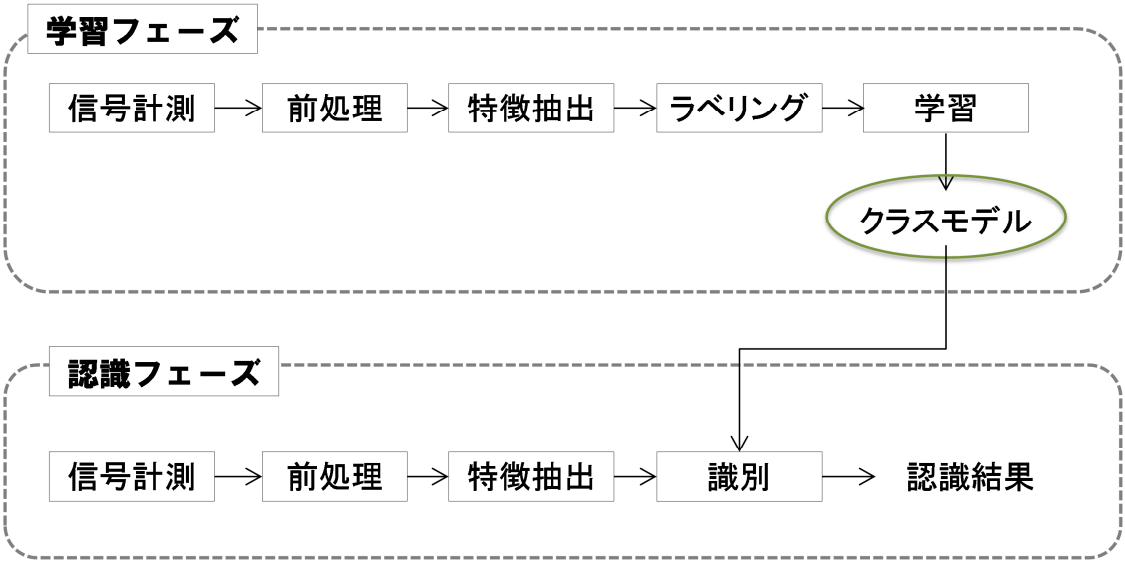

図 3.1: 筋電インタフェースの処理の流れ

まず，信号計測部において顔や腕などの目的箇所に電極を装着して筋電位信号を計測し，増幅や整流･平滑化などの前処理を行う．そして次に，前処理済みの筋電位から特徴を抽出し，同時刻に得られた各電極からの特徴を合わせて一つの特徴ベクトルとする．学習部では，この特徴ベクトルが得られた際にどのような動作が行われていたかというラベルを各特徴ベクトルに与え，あらかじめモデルを学習しておく．認識の際には，学習したモデルを基に特徴ベクトルを認識し，現在行われていると推定される動作のラベルを出力する．次節から，各処理について詳しく解説する．

3.2　前処理

筋電位の計測を行う際，図 3.2 に示す二種類の信号を各電極で得る．一つは，計測した生の筋電位信号に対して，ノイズ除去及び増幅の処理を行ったもので，以降ではこの信号を EMG（electromyogram）と呼ぶ．もう一つは，EMG を全波整流・平滑化した信号で，これを IEMG（積分筋電位）と呼ぶ．運動生理学分野においては，IEMG の振幅が筋活動の指標としてよく用いられる．

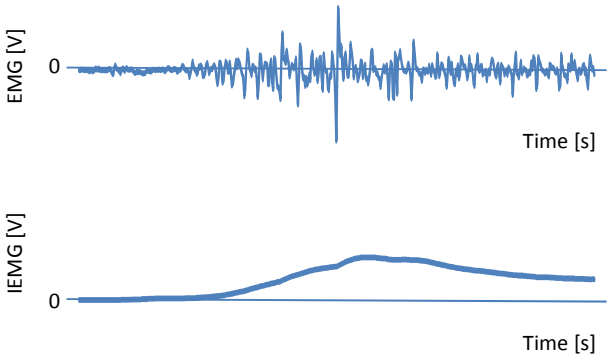

図 3.2: EMG と IEMG の波形

3.3　特徴抽出

筋電位のような時系列信号の特徴を抽出する際には，フレーム化と呼ばれる処理が行われることが多い．これは，入力信号が短時間区間において定常確率過程に従うと仮定し，その信号区間を切り出す処理のことで，特徴抽出ではこのフレームごとに一つの特徴ベクトルを抽出することになる．

3.3.1　時間領域での特徴量

フレーム内平均 IEMG

第 p フレームでのフレーム内平均 IEMG は $AIEMG(p)$ であらわすこととし，以下の式で求められる．

$$AIEMG(p) = \frac{1}{N}\sum_{n=0}^{N-1} I(n) \tag{3.1}$$

ここで，N は第 p フレーム内のサンプル数，$I(n)$ は第 p フレーム内で n 点目の IEMG サンプルを表す．この特徴は信号の平均振幅を示すもので，筋電位における特徴で最も重要な特徴の一つである [23].

3.3.2　周波数領域での特徴量

ケプストラム係数

信号の周波数成分の解析手法として，短時間フーリエ変換（Short-Time Fourier Transform :STFT）がよく知られる．STFT はフレーム内の信号に対して離散フーリエ変換（Discrete Fourier Transform:DFT）を行うもので，EMG の DFT は次のように表される．

$$X_p(k) = \sum_{n=0}^{N-1} E(n)\ e^{-j2\pi kn/N} \tag{3.2}$$

$E(n)$は第pフレーム内でn点目のEMGサンプルを表す．フーリエ変換では，与えられた信号を周期信号と仮定するため，フレーム内の最初の値と最後の値が不連続になる場合，スペクトルに誤差が生じる．そのため，信号に窓関数をかけて最初と最後の振幅を小さくする．そしてこれを用いて，ケプストラム係数$CC_n(p)$は次式で計算される．

$$CC_p(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log|X_p(k)|\ e^{j2\pi kn/N} \tag{3.3}$$

ケプストラム分析を行うことにより，ケプストラム係数の低次項にパワースペクトルの包絡形状の特徴，高次項にはパワースペクトルの微細構造の特徴が分離して得られる．音声認識において，ケプストラム分析は有効な分析手法としてよく用いられるが，この場合には低次項は調音特性，高次項は音源特性を示す．

第4章　Hybrid HMM/SVMによる認識モデル

4.1　隠れマルコフモデル（HMM）

HMMは，出力のシンボル系列からは一意に状態遷移系列を決定することができない非決定性確率有限状態オートマトン[†1]に対応付けることができる．HMMは，シンボル$\boldsymbol{x}_t$を出力する確率分布が$b_j(\boldsymbol{x}_t)$であるような信号源(状態)が状態遷移確率a_{ij}をもって接続されたものとして定義される．ただし，a_{ij}は状態q_iからq_jに遷移する確率$P(q_t = j|q_{t-1} = i)$を表す．つまり，次状態は現状態のみに依存して確率的に定まり，現状態とそこからの出力にも確率的な依存関係がある．HMMは形状と出力確率分布によって表4.1のように大別される．

表 4.1: HMM の分類

形状による分類	Ergodic HMM
	Left-to-Right HMM
出力分布による分類	離散型 HMM
	連続型 HMM

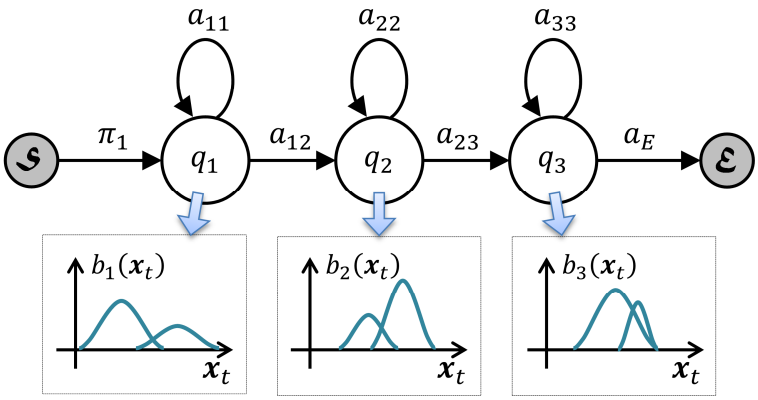

図 4.1: Left-to-Right HMM

Ergodic HMMは，すべての状態間の遷移が可能であり，次の状態に遷移しても前の状態に戻ることが可能なモデルである．Left-to-Right HMMは，遷移が一方向に進むモデルであり，次の状態に遷移すると前の状態には戻ることができない．音声認識では，音声信号にエルゴード性[†2]がないという仮定からLeft-to-Right HMMが利用される．また，出力が音

[†1] オートマトンとは，入力に従って内部状態を変化させ，入力列が受理できるかどうかを判定する仮想機械
[†2] 集合平均と時間平均が一致することをエルゴード性が成立するという

声信号の特徴ベクトルという連続量に対応するため，出力確率分布は連続出力分布として式 (4.1) で表される混合正規分布が用いられることが多い．

$$\begin{aligned} b_j(\boldsymbol{x}_t) &= \sum_{m=1}^{M_j} c_{jm}\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_{jm}) \\ &= \sum_{m=1}^{M_j} c_{jm}\frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{x}_t-\boldsymbol{\mu}_{jm}\right)^{\top}\boldsymbol{\Sigma}_{jm}^{-1}\left(\boldsymbol{x}_t-\boldsymbol{\mu}_{jm}\right)\right)}{\sqrt{(2\pi)^N\,|\boldsymbol{\Sigma}_{jm}|}} \end{aligned} \tag{4.1}$$

ただし，N は $\boldsymbol{x}$ の次元数，M_j は状態 q_j における正規分布の混合数，c_{jm} は混合重み，$\boldsymbol{\mu}_{jm}$ と $\boldsymbol{\Sigma}_{jm}$ はそれぞれ正規分布の平均値ベクトルと共分散行列を表す．

HMM$\mathcal{M}$ のモデルパラメータ w は，状態数を S とすると，初期状態確率 $\pi=\{\pi_i\}_{i=1}^{S}$，状態遷移確率 $A=\{a_{ij}\}_{i,j=1}^{S}$，各状態 i での出力確率 $B=\{b_i(\boldsymbol{x}_t)\}_{i=1}^{S}$ により，$w=(A,B,\pi)$ として与えられる．このとき，状態が $Q=\{q_1,q_2,...,q_T\}$ と遷移して，観測ベクトル系列 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_T)$ が出力される同時確率 $P(\boldsymbol{X},Q|w)$ は，$P(\boldsymbol{X}|Q,w)$ と $P(Q|w)$ の同時確率で，

$$\begin{aligned} P(\boldsymbol{X},Q|w) &= P(\boldsymbol{X}|Q,w)\;P(Q|w) \\ &= \prod_{t=1}^{T} a_{q_{t-1}q_t}b_{q_t}(\boldsymbol{x}_t) \end{aligned} \tag{4.2}$$

となる．ただし，$a_{0i}=\pi_i$ とする．したがって，観測ベクトル系列 $\boldsymbol{X}$ が HMM$\mathcal{M}$ から出力される確率は，すべての可能な状態遷移の組み合わせについて，和をとることにより求められる．

$$\begin{aligned} P(\boldsymbol{X}|w) &= \sum_{Q} P(\boldsymbol{X},Q|w) \\ &= \sum_{Q}\prod_{t=1}^{T} a_{q_{t-1}q_t}b_{q_t}(\boldsymbol{x}_t) \end{aligned} \tag{4.3}$$

4.2　Baum-Welch アルゴリズムによる HMM の学習

HMM の学習は，観測データ $\{\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},...,\boldsymbol{X}^{(l)}\}$ に対して，式 (4.3) で与えられる観測尤度 $P(\boldsymbol{X}|w)$ を最大にするモデルパラメータ $w=(A,B,\pi)$ を求めることである．すなわち，

$$w_{\max}=\arg\max_{w} P(\boldsymbol{X}|w) \tag{4.4}$$

しかし，HMM の状態 Q と観測ベクトル系列 $\boldsymbol{X}$ の対応関係は隠れ変数となって一意には決まらないため，式 (4.4) を解析的に導くことはできない．そこで，推定パラメータに初期値を与えたモデルを，観測データを利用することで改善していくという方法をとる．具体的には，Expectation-Maximization アルゴリズムを HMM のパラメータ再推定に適用した Bauch-Welch アルゴリズムにより，局所最適解の推定が行われる．

Forward アルゴリズム

式 (4.3) は，観測ベクトル系列の長さ T，HMM の状態数 S のとき，$\mathcal{O}(2TS^T)$ の計算量が必要になる．そこで，計算量の削減のために Forward アルゴリズムが考案されされており，計算量を $\mathcal{O}(TS^2)$ に抑えることができる．これは，図 4.2 に示すように，HMM の各状態と各時刻の観測ベクトル系列を格子状に配置したトレリスを基にしたもので，可能な状態遷移系列をすべて計算するのではなく，格子点に至るまでのすべての状態遷移系列に関する計算結果を記憶しておき，その結果を再利用することにより，計算量の削減が可能となる．

時刻 $t-1$ までに観測ベクトル系列 $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{t-1} (t \leq T)$ を出力し，時刻 t に状態 q_i から q_j に遷移して $\boldsymbol{x}_t$ を出力する確率として前向き確率 $\alpha_t(j)$ を定義する．$\alpha_t(j)$ は

$$\alpha_t(j) = \begin{cases} \pi_j & (t=1) \\ \sum_{i=1}^{S} \alpha_{t-1}(i) a_{ij} b_j(\boldsymbol{x}_t) & (t>1) \end{cases} \tag{4.5}$$

のように漸化的に求めることができ，HMM $\mathcal{M}$ から $\boldsymbol{X}$ が出力される確率は式 (4.6) で求められる．

$$P(\boldsymbol{X}|w) = \sum_{j \in \mathcal{F}} \alpha_T(j) \tag{4.6}$$

ただし，$\mathcal{F}$ は終了状態の集合を表す.

Backward アルゴリズム

Baum-Welch アルゴリズムでは，時刻 t に q_i（この状態を以降では $g_t(i)$ として表す[†3]）を通過する事後確率，及び $g_t(i)$ を通過後に $g_{t+1}(j)$ に遷移する事後確率が必要となる．そこで，$g_t(i)$ を通過後に $g_{t+1}(j)$ に遷移し，さらにそこから観測ベクトル系列 $\boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+2}, ..., \boldsymbol{x}_T$ を出力する確率を後向き確率 $\beta_t(i)$ として定義する．このための処理は Forward アルゴリズムと逆向きの処理となるため Backward アルゴリズムと呼ばれる．

$$\beta_t(i) = \begin{cases} [q_i \in \mathcal{F}] & (t=T) \\ \sum_{j=1}^{S} a_{ij} b_j(\boldsymbol{x}_{t+1}) \beta_{t+1}(i) & (t<T) \end{cases} \tag{4.7}$$

ただし，$[\mathcal{P}]$ の表記は

$$[\mathcal{P}] = \begin{cases} 1 & \text{if } \mathcal{P} \text{ is true,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

を示すものとする [29].

Baum-Welch アルゴリズム

観測ベクトル系列 $\boldsymbol{X}$ が与えられたときに，$g_t(i)$ を通過する事後確率を $\gamma_t(i)$，$g_t(i)$ を通過して $g_{t+1}(j)$ に遷移する事後確率を $\xi_t(i,j)$ とする．これらを求めることが Expectation ス

[†3] 図 4.2 の時刻 t，状態 q_i における格子点に対応する

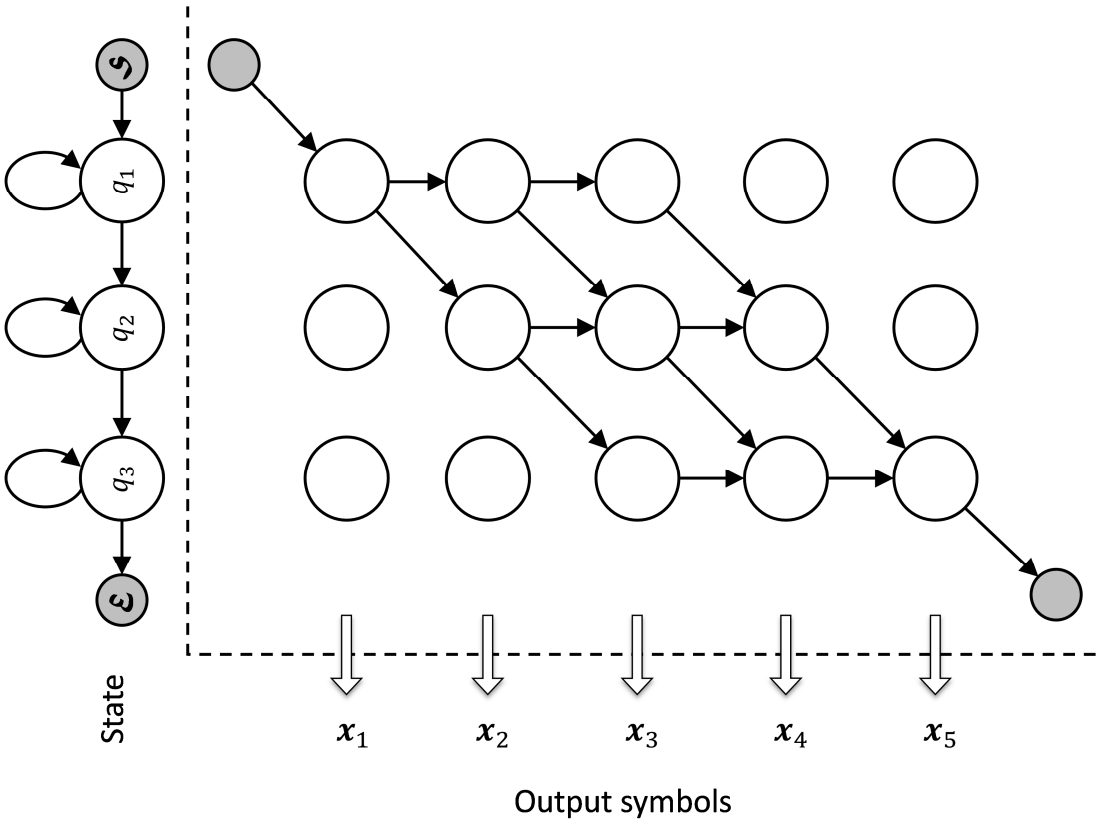

図 4.2: Left-to-Right HMM のトレリス表現

テップに相当する．Baum-Welch アルゴリズムでは，$\gamma_t(i)$，$\xi_t(i,j)$ に基づいて HMM のパラメータである初期状態確率 π_i，遷移確率 a_{ij}，各状態での出力確率 $b_i(\boldsymbol{x}_t)$ を再推定していく．

$$\begin{aligned}\gamma_t(i) &= \frac{P(q_t=i,\boldsymbol{X}|w)}{\sum_{i=1}^{S}P(q_t=i,\boldsymbol{X}|w)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{S}\alpha_t(i)\beta_t(i)}\end{aligned} \tag{4.9}$$

$$\begin{aligned}\xi_t(i,j) &= \frac{P(q_t=i,q_{t+1}=j|\boldsymbol{X},w)}{\sum_{i=1}^{S}\sum_{j=1}^{S}P(q_t=i,q_{t+1}=j|\boldsymbol{X},w)} \\ &= \frac{\alpha_t(i)a_{ij}b_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{S}\sum_{j=1}^{S}\alpha_t(i)a_{ij}b_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j)}\end{aligned} \tag{4.10}$$

Maximization ステップでは，Expectation ステップで求めた値を利用してパラメータの再推定を行う．再推定したパラメータ $\hat{\pi}_i$，$\hat{a}_{ij}$，$\hat{b}_i(\boldsymbol{\mathcal{K}})$ は，

$$\hat{\pi}_i = \gamma_1(i) \tag{4.11}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{4.12}$$

$$\hat{b}_i(\boldsymbol{\mathcal{K}}) = \frac{\sum_{t=1}^{T}[\boldsymbol{x}_t=\boldsymbol{\mathcal{K}}]\gamma_t(i,j)}{\sum_{t=1}^{T}\gamma_t(i)} \tag{4.13}$$

と表せる．このパラメータ $\hat{w}$ を用いてモデル $\mathcal{M}$ を更新すると，$P(\boldsymbol{X}|\hat{w}) > P(\boldsymbol{X}|w)$ を満たすことが示されており [30]，$\hat{w}$ をまた w として繰り返し再推定を行うと，$\hat{w}$ は局所最適解に収束する．

ここではさらに，出力確率分布を式 (4.1) で表される混合正規分布とする場合について述べる．各パラメータは，式 (4.13)，及び，以下に示す HMM のパラメータの制約条件を用いて，ラグランジュの未定乗数法によって求めることができる．

1. 存在するすべての初期状態確率の総和は 1.0 になる．

$$\sum_{i=1}^{S}\pi_i = 1.0 \tag{4.14}$$

2. ある状態 q_i から遷移する可能性のあるすべての状態 q_j への状態遷移確率の総和は 1.0 になる．

$$\sum_{j=1}^{S}a_{ij} = 1.0 \tag{4.15}$$

3. ある状態 q_i において出力される可能性のあるすべてのシンボル系列の出力確率の総和は 1.0 になる．

$$\sum_{\boldsymbol{\mathcal{K}}}b_i(\boldsymbol{\mathcal{K}}) = 1.0 \tag{4.16}$$

再推定したパラメータ $\hat{\boldsymbol{\mu}}_{jm}$，$\hat{\boldsymbol{\Sigma}}_{jm}$，$\hat{c}_{jm}$ は，

$$\hat{\boldsymbol{\mu}}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)\, \boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_t(i,m)} \tag{4.17}$$

$$\hat{\boldsymbol{\Sigma}}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)\, (\boldsymbol{x}_t - \boldsymbol{\mu}_{im})(\boldsymbol{x}_t - \boldsymbol{\mu}_{im})^{\top}}{\sum_{t=1}^{T} \gamma_t(i,m)} \tag{4.18}$$

$$\hat{c}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{4.19}$$

ただし $\gamma_t(i,m)$ は，$\boldsymbol{x}_t$ が観測されたときに，$g_t(i)$ を通過して HMM の状態 q_i における m 番目の正規分布から出力される事後確率とし，

$$\gamma_t(i,m) = \gamma_t(i)\, \frac{c_{im}\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im})}{\sum_{m=1}^{M_i} c_{im}\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im})} \tag{4.20}$$

とする．

4.3　Viterbi アルゴリズムによる認識

HMM を利用した認識とは，観測された時系列データを最も高い確率で出力する HMM を見つけることである．つまり，HMM$\mathcal{M}$ のパラメータ w が与えられたときに，観測系列 $\boldsymbol{X}$ を出力する確率 $P(\boldsymbol{X}|w)$ を求める必要があるが，Viterbi アルゴリズムによって近似的に効率よく計算することができる．前節で述べた Forward アルゴリズムがすべての状態遷移系列の確率の和を計算しているのに対して，Viterbi アルゴリズムでは最大確率を与える状態遷移系列のみを考慮する．前向き確率 $\alpha'_t(j)$ を次のように定義する．

$$\alpha'_t(j) = \begin{cases} \pi_j & (t=0) \\ \max_i \alpha'_{t-1}(i) a_{ij} b_j(\boldsymbol{x}_t) & (t \geq 1) \end{cases} \tag{4.21}$$

このとき，確率の積を逐次計算していくと最終的にアンダーフローが生じることがあるため，対数尤度が利用されることが多い．式 (4.21) より，$P(\boldsymbol{X}|w)$ は，

$$P(\boldsymbol{X}|w) = \max_{j \in \mathcal{F}} \alpha'_T(j) \tag{4.22}$$

として求めることができる．また，Viterbi アルゴリズムではバックトラックにより，最大確率を与える状態遷移系列 $Q_{\max}$ を得ることができ[†4]，観測系列 $\boldsymbol{X}$ を $Q_{\max}$ の各状態ごとに分割することを Viterbi アラインメントと呼ぶ．Viterbi アルゴリズムの計算量は最悪でも $O(ST)$ で収まる．

音声認識では，認識対象とする単語ごとに HMM を $w_1, w_2, ..., w_U$ と作成しておき，観測特徴ベクトル系列 $\boldsymbol{X}$ に対して，

$$w_{u_{\max}} = \arg\min_{w_u} P(\boldsymbol{X}|w_u) P(w_u) \tag{4.23}$$

という $w_{u_{\max}}$ を求めることになる．

[†4] これを Viterbi 経路と呼ぶ

4.4　サポートベクトルマシン（SVM）

SVMは教師あり学習による認識手法の一つであり，マージン最大化基準により学習を行うという特徴を持つ．

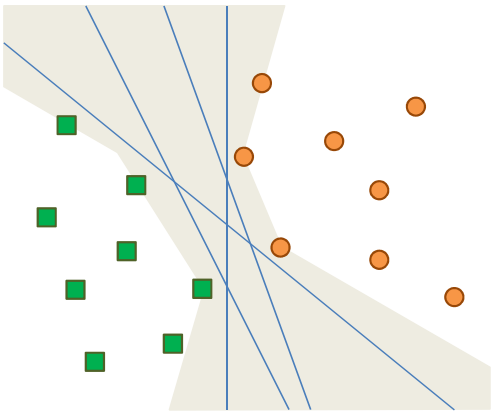

図 4.3: 線形分離可能なデータと決定境界

図 4.3 のようなデータをクラス分けすることを考えたとき，すべてのデータを誤りなくクラス分けすることのできる境界は無数に存在する．SVM ではマージン最大化に基づき，図 4.4 のように，決定境界とそれに最も近い位置にあるサンプルとの距離（マージン）が最大となるように境界を決定する[†5]．つまり，各クラスからのサポートベクトルの垂直二等分超平面が決定境界となる．

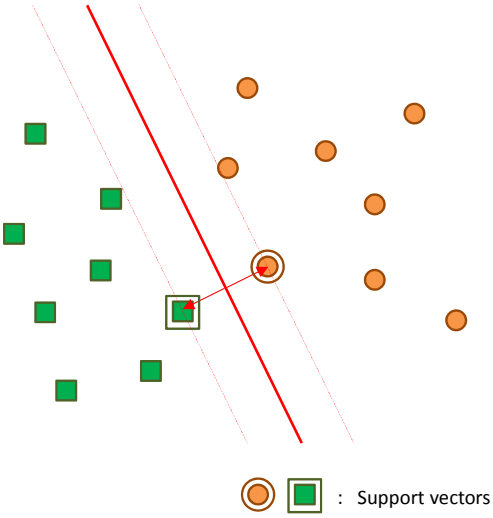

図 4.4: マージン最大化基準での決定境界

[†5] 決定境界に最も近い位置にあるサンプルをサポートベクトル（Support Vector）という

4.5　SVM の学習アルゴリズムと識別関数

線形 SVM

学習データ $\boldsymbol{X}$ を次のように定義する．

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_l\} \tag{4.24}$$

今，2 クラスが線形分離可能な問題について考え，式 (4.24) の各サンプルに対応するクラスラベルを

$$\{y_1, y_2, ..., y_l\} \quad (y_i \in \{1, -1\}) \tag{4.25}$$

とし，クラス間の境界を示す超平面を次のように表す.

$$g(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b = 0 \tag{4.26}$$

すると，学習サンプル $\boldsymbol{x}$ と決定境界 $g(\boldsymbol{x})$ との距離は

$$\frac{|\boldsymbol{w}^\top \boldsymbol{x} + b|}{||\boldsymbol{w}||} \tag{4.27}$$

となり，サポートベクトルと決定境界との距離は次のように表せる．

$$\min_i \frac{|\boldsymbol{w}^\top \boldsymbol{x}_i + b|}{||\boldsymbol{w}||} \tag{4.28}$$

ここで，式 (4.26) において，$\boldsymbol{w}$ や b を定数倍しても超平面そのものは変化しないことから，

$$\min_i |\boldsymbol{w}^\top \boldsymbol{x}_i + b| = 1 \tag{4.29}$$

という制約を設けることができ，これにより式 (4.28) は次のように書ける．

$$\frac{1}{||\boldsymbol{w}||} \tag{4.30}$$

以上より，マージンを最大化するという問題は

$$\begin{aligned} &\text{minimize}_{\boldsymbol{w},b} && G(\boldsymbol{w}) = ||\boldsymbol{w}||^2, \\ &\text{subject to} && y_i(\boldsymbol{w}^\top \boldsymbol{x} + b) \geq 1 \quad (i = 1, ..., l) \end{aligned} \tag{4.31}$$

という最小化問題として表現できる．すなわち，SVM における学習とは，式 (4.26) のパラメータとして，上の問題の最適解 $\boldsymbol{w}^*$ と b^* を求めるということになる．

この最適化問題のラグランジアン $\mathcal{L}$ は，ラグランジュ乗数 $\lambda_i > 0$ を導入して，

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{l} \lambda_i \left(y_i \left(\boldsymbol{w}^\top \boldsymbol{x} + b \right) - 1 \right) \tag{4.32}$$

となる．そして，$\mathcal{L}$ の勾配に関して．定常性の仮定から以下の関係を得る．

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \quad \Longrightarrow \quad \boldsymbol{w} - \sum_{i=1}^{l} \lambda_i y_i \boldsymbol{x}_i = 0 \tag{4.33}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{l} \lambda_i y_i = 0 \tag{4.34}$$

これらを式 (4.32) に代入して，式 (4.32) は次の双対問題に帰着する．

$$\begin{aligned} &\text{maxmize}_{\boldsymbol{\lambda}} \qquad \mathcal{W}(\boldsymbol{\lambda}) = \sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^{\top} \boldsymbol{x}_j, \\ &\text{subject to} \qquad \sum_{i=1}^{l} \lambda_i y_i = 0,\ \alpha_i \geq 0 \quad (i = 1, ..., l) \end{aligned} \tag{4.35}$$

これを解くと多くの α_i が 0 となり，式 (4.33) から，

$$\boldsymbol{w}^* = \sum_{i=1}^{l} \lambda_i^* y_i \boldsymbol{x}_i \tag{4.36}$$

で最適なパラメータを決定することができる．$\boldsymbol{w}^*$ は，$\alpha_i \neq 0$ と対応する学習サンプル $\boldsymbol{x}_i$ のみから決まり，式 (4.31) 及び (4.35) の最適解 $\boldsymbol{w}^*$ と b^*，$\boldsymbol{\lambda}^*$ に関して，カルッシュ - クーン - タッカー相補条件

$$\boldsymbol{\lambda}^* \left(y_i \left(\boldsymbol{w}^{\top} \boldsymbol{x} + b \right) - 1 \right) = 0 \tag{4.37}$$

が成立することから，この $\boldsymbol{x}_i$ はサポートベクトルである．b の値は双対問題には現れないので b^* は主問題の制約を利用して発見することになり，各クラスに属する任意のサポートベクトル $\boldsymbol{x}_{sv:1}$ および $\boldsymbol{x}_{sv:-1}$ を用いて

$$b^* = -\frac{1}{2} \left(\boldsymbol{w}^{*\top} \boldsymbol{x}_{sv:1} + \boldsymbol{w}^{*\top} \boldsymbol{x}_{sv:-1} \right) \tag{4.38}$$

で求められる．最終的に，線形 SVM による識別関数 $f(\boldsymbol{x})$ は，

$$\begin{aligned} f(\boldsymbol{x}) &= \text{sign}(\boldsymbol{w}^{*\top} \boldsymbol{x} + b^*) \\ &= \text{sign} \left(\sum_{i=1}^{l} \lambda_i^* y_i \boldsymbol{x}_i^{\top} \boldsymbol{x} + b^* \right) \end{aligned} \tag{4.39}$$

ソフトマージン

図 4.5 に示すように，2 クラスが線形分離不可能な場合，前述の解法では解なしとなって決定境界が求まらない．そこで，スラック変数 $\xi_i (i = i, ..., l)$ を導入し，学習データに対して若干の誤りを許すように条件を緩め，

$$\begin{aligned} &\text{minimize}_{\boldsymbol{w}, b, \boldsymbol{\xi}} \qquad G(\boldsymbol{w}, \boldsymbol{\xi}) = ||\boldsymbol{w}||^2 + C \sum_{i=1}^{l} \xi_i, \\ &\text{subject to} \qquad y_i (\boldsymbol{w}^{\top} \boldsymbol{x} + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (i = 1, ..., l) \end{aligned} \tag{4.40}$$

として問題を定める．ここで，第 2 項はマージンからはみ出した学習サンプルに対するペナルティ項であり，パラメータ C は第 1 項と第 2 項のバランスを決める値で，問題に応じて適切に与える．

式 (4.40) に対応する双対問題は次のようになる．

$$\begin{aligned} &\text{maxmize}_{\boldsymbol{\lambda}} && \mathcal{W}(\boldsymbol{\lambda}) = \sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j, \\ &\text{subject to} && \sum_{i=1}^{l} \lambda_i y_i = 0,\ 0 \leq \alpha_i \leq C \quad (i = 1, ..., l) \end{aligned} \tag{4.41}$$

この凸 2 次計画問題の最適解 $\boldsymbol{w}^*$ は勾配法を用いて求めることができる．ソフトマージン線形 SVM の識別関数は，最終的に式 (4.39) と同様である．

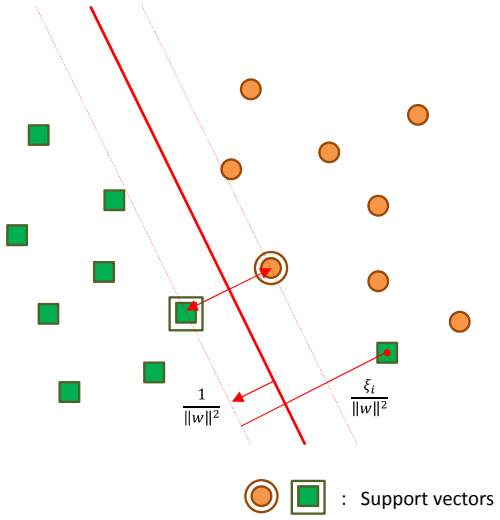

図 4.5: 線形分離不可能なデータ分布

非線形 SVM

ソフトマージンを用いたとしても，学習データが本質的に非線形で複雑な問題の場合，非線形 SVM を用いる．非線形 SVM では，線形分離不可能な学習データを高次元の特徴空間に写像し，写像先の特徴空間において線形分離可能な決定境界を求める．ここで，次元 N のパターン $\boldsymbol{x}$ を N' 次元空間に写像する関数を $\Phi(\boldsymbol{x})$ とおくと，識別関数は

$$f(\boldsymbol{x}) \quad = \quad \text{sign}(\boldsymbol{w}^\top \Phi(\boldsymbol{x}) + b) \tag{4.42}$$

と表せる．さらに最小化問題は，非線形 SVM では次のような双対問題となる．

$$\begin{aligned} &\text{maxmize}_{\boldsymbol{\lambda}} && \mathcal{W}(\boldsymbol{\lambda}) = \sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_j), \\ &\text{subject to} && \sum_{i=1}^{l} \lambda_i y_i = 0,\ 0 \leq \alpha_i \leq C \quad (i = 1, ..., l) \end{aligned} \tag{4.43}$$

式 (4.43) の内積計算 $\Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_j)$ をそのまま計算しようとすると，写像された高次元空間でのベクトル演算となり，計算量が膨大になってしまう．そこで，Φ が次を満たすようなカーネル関数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ を用意する．

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_j) \tag{4.44}$$

式 (4.43) で，カーネル関数を用いて内積計算を置き換えることにより，

$$\begin{aligned} &\text{maxmize}_{\boldsymbol{\lambda}} \qquad \mathcal{W}(\boldsymbol{\lambda}) = \sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \\ &\text{subject to} \qquad \sum_{i=1}^{l} \lambda_i y_i = 0,\ 0 \leq \alpha_i \leq C \quad (i = 1, ..., l) \end{aligned} \tag{4.45}$$

となる．このように高次元空間に写像しながら，写像された空間での内積計算をカーネル関数で置き換えて実際には避ける方法は，カーネルトリックと呼ばれる．カーネル関数としては，式 (4.46) で定義される多項式カーネルや，式 (4.47) の動径基底関数（Radial Basis Function:RBF）などが知られている．

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j + 1)^d \tag{4.46}$$

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2) \tag{4.47}$$

最終的な非線形 SVM の識別関数は

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{l} \lambda_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^* \right) \tag{4.48}$$

となり，最適解 λ_i^* を効率よく求めるためのアルゴリズムとして，逐次最小最適化（Sequential Minimal Optimisation:SMO）アルゴリズム [32] がある．

多クラス問題への SVM の拡張

SVM は原理的に 2 クラス問題に対応した認識器であるため，多クラスの認識問題は，2 クラスの認識問題の組み合わせとして表現する．つまり，Y 個のクラスのすべての組み合わせで $Y(Y+1)/2$ 個の決定境界を求め，各決定境界により認識を行う．そして，すべての認識結果による投票処理により，最も多く認識されたクラスを最終的な結果とする．

4.6　Hybrid HMM/SVM の概要

前節で述べたとおり，SVM はそのマージン最大化基準によって定めた識別関数によって高い汎化性能をもち，多くの問題に対して優れた認識性能をもつ．また，比較的学習データの数が少ない場合でも高い認識能力が期待できる．しかし，時系列データを扱うような問題においては，そのパターン変化を捉えることができないという問題がある．一方，HMM は複数の定常状態の遷移によって時系列パターンのモデル化を行うことで，音声認識への

応用において優れた性能を示している．特に，状態間の遷移確率と各状態での出力確率により，確率的な表現を行うため，確率モデルで表現される言語レベルの処理と統合しやすいといった点も注目できる．しかし，統計的手法であるためにパラメータの推定には多量の学習サンプルが必要であることが知られる．そこで，SVM のもつ高い認識性能を利用しながら，時系列パターンの伸縮を表現できるようにモデル化を行うことを考える．すなわち，HMM の構造を取り入れて SVM を連鎖させることで，時間的な状態遷移を表現する．そして，学習サンプルと連鎖モデルの各状態とを対応付けるために，HMM を利用した Viterbi アラインメントを利用する．Viterbi アラインメントでは，各状態に分割された区間のデータは定常的であることが期待できる．そのため，この各状態ラベルをクラスラベルとして SVM の学習を行うことで，SVM の弁別的な認識で有効だと考えられる．また，各状態間を遷移する頻度を遷移確率とし，これによって各状態及び状態を連結させたパターンの時間伸縮にも対応することが可能となる．さらに，確率モデルとして記述するため，各状態での出力確率には，次に述べる SVM による事後確率を用いることとする．これによって，図 4.6 に示すように HMM の各状態での出力確率を SVM による事後確率に置き換えたような構成となるため，認識の際も HMM と同様に Viterbi アルゴリズムを利用する．

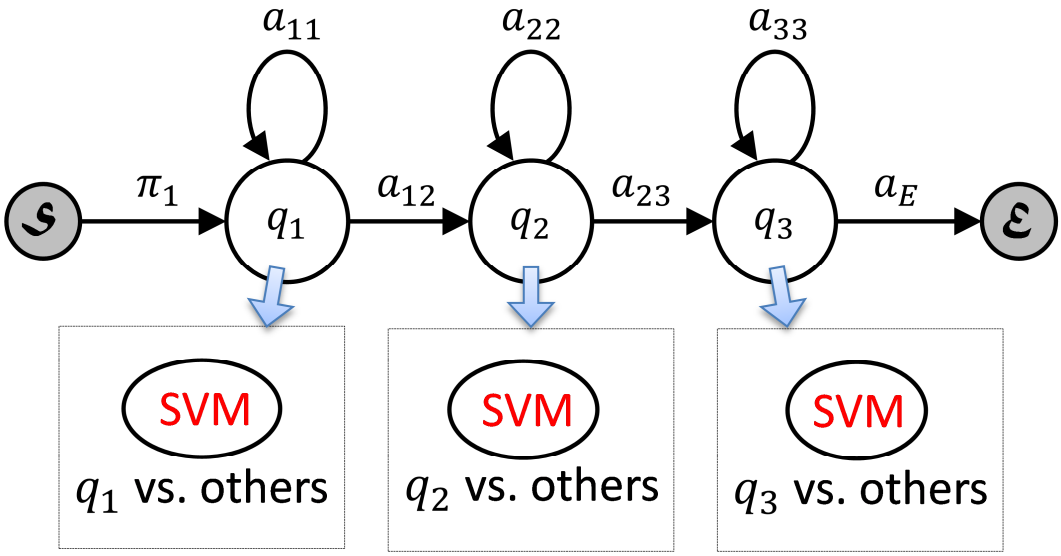

図 4.6: Hybrid HMM/SVM の構造イメージ

SVM による事後確率

SVMから事後確率を導出するために，マージン最大化基準で定めた決定境界からの距離によって確率を求めることを考える．すなわち，決定境界から近い距離にあるデータに対しては小さな確率を与え，決定境界から遠い場合には，大きな確率を与えるようにする．これを実現するために，次のようなシグモイド関数を用意する（図 4.7）．

$$P(y=1|\boldsymbol{x}) = \frac{1}{1+\exp(Ag(\boldsymbol{x})+B)} \tag{4.49}$$

このパラメータ A，B を，学習データを用いて次のように求める．

$$\arg\min_{A,B} \quad -\sum_i t_i \log(p_i) + (1-t_i)\log(1-p_i) \tag{4.50}$$

ただし，

$$t_i = \frac{y_i + 1}{2} \tag{4.51}$$

$$p_i = \frac{1}{1+\exp(Ag(\boldsymbol{x}_i)+B)} \tag{4.52}$$

$$g(\boldsymbol{x}) = \sum_{i=1}^{l} \lambda_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^* \tag{4.53}$$

これによって，SVM による事後確率を近似できることが実験的に示されている [33]．二次元平面上で 3 クラス（青，赤，緑）に属する各点を学習データとして与えたときの決定境界，及び平面上の各点の青クラスの事後確率を示したものを例として図 4.8 に示す．

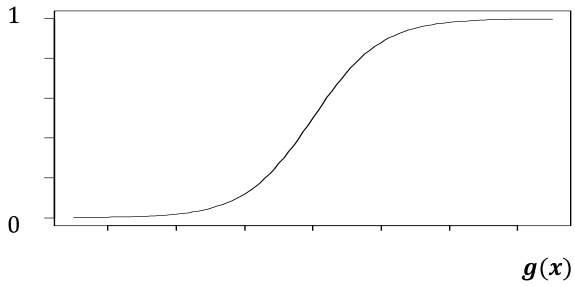

図 4.7: シグモイド関数

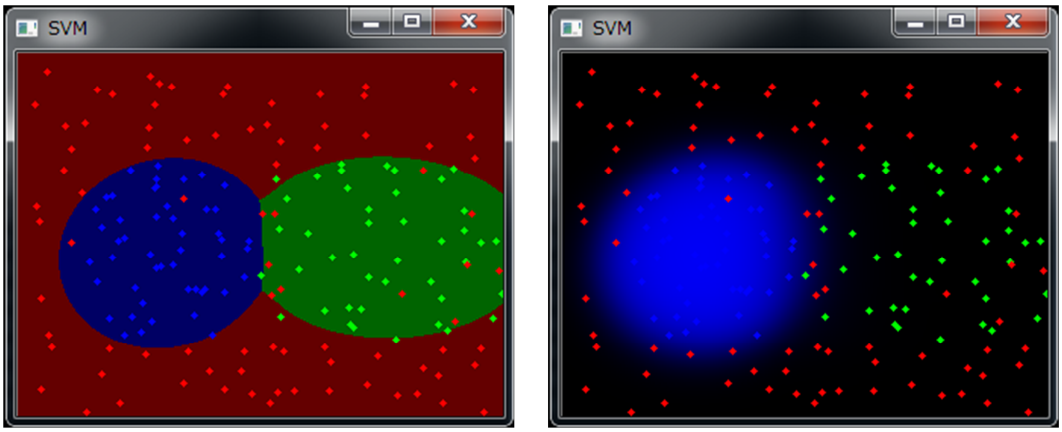

図 4.8: 二次元データの SVM3 クラス問題における決定境界と事後確率の様子

4.7　Hybrid HMM/SVM による認識モデルの生成

本研究で用いる筋電インタフェースのシステムを，Hybrid HMM/SVM を利用して構築することを考える．学習フェーズの処理の流れを図 4.9 に示す．

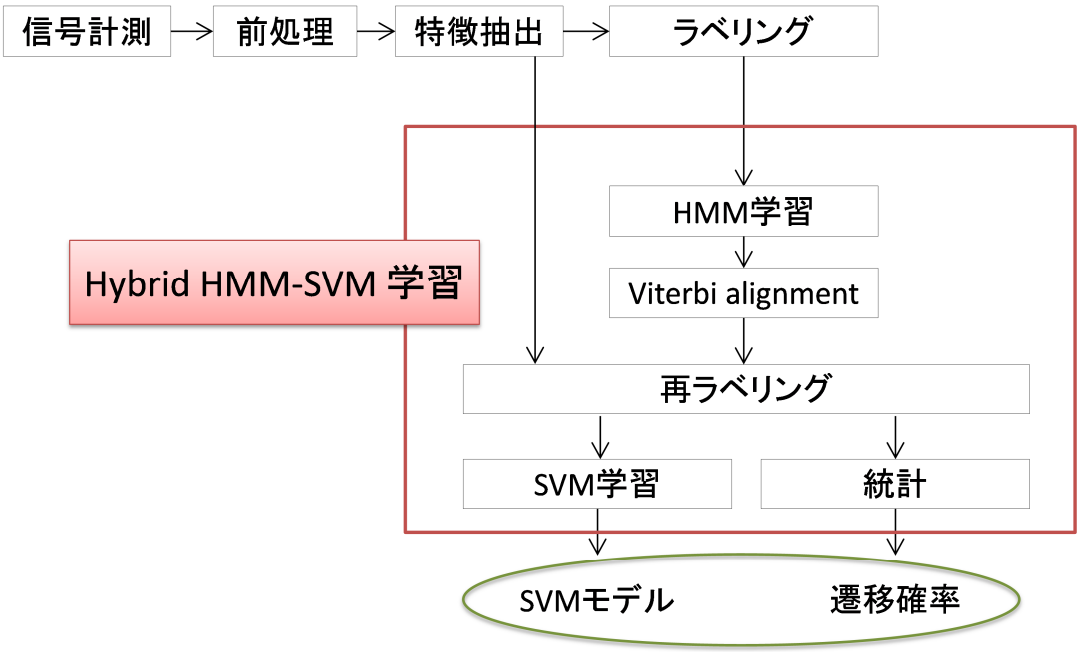

図 4.9: 学習フェーズの処理

Hybrid HMM/SVM では，ある動作に関するモデルを作成する場合に，その動作の時系列でのパターンを各状態区間に分割する．しかし，時間伸縮があるパターンに対して正解ラベルを用意しておくことは難しい．そこで，各動作ラベルを正解ラベルとして，まず HMM 学習を行う．そして，学習データで Viterbi アラインメントを行うことで，図 4.10 に示すように状態系列を得る．これによる再ラベリングの例を図 4.11 に示す．正解ラベルとして「あ」をつけていたデータに対して，3 状態 HMM により Viterbi アラインメントを行うことにより,「あ」の信号の立ち上がり部分と「あ」の定常区間，そして立下りの部分というような詳細な構造を得ることができる．

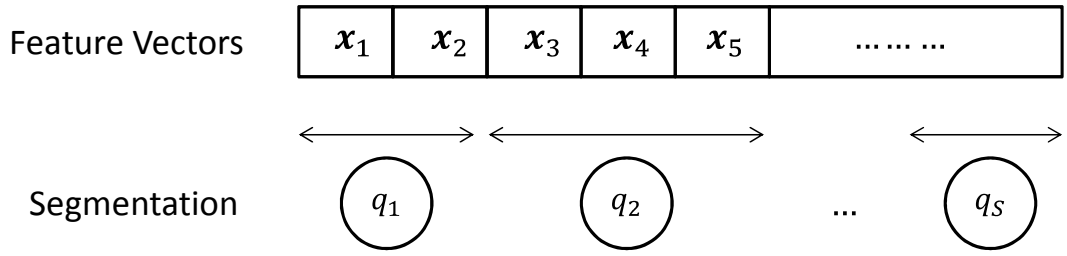

図 4.10: Viterbi アラインメント

状態系列の前後の出現頻度を数えることで遷移確率を求める．SVM 学習では,（認識対象の動作数）×（状態数）個のクラスでの認識となる．Hybrid HMM/SVM による認識では，対

象の動作についてそれぞれモデルを作成しておく必要がある．また，モデルを連結していくことで，一連の動作の認識を行うことも可能である．黙声認識に適用する場合には，母音や子音単位でモデルを作成しておき，それらを連結して単語などに拡張していくことになる．連続黙声認識では，音声認識と同様に言語モデルを用いることが可能である．

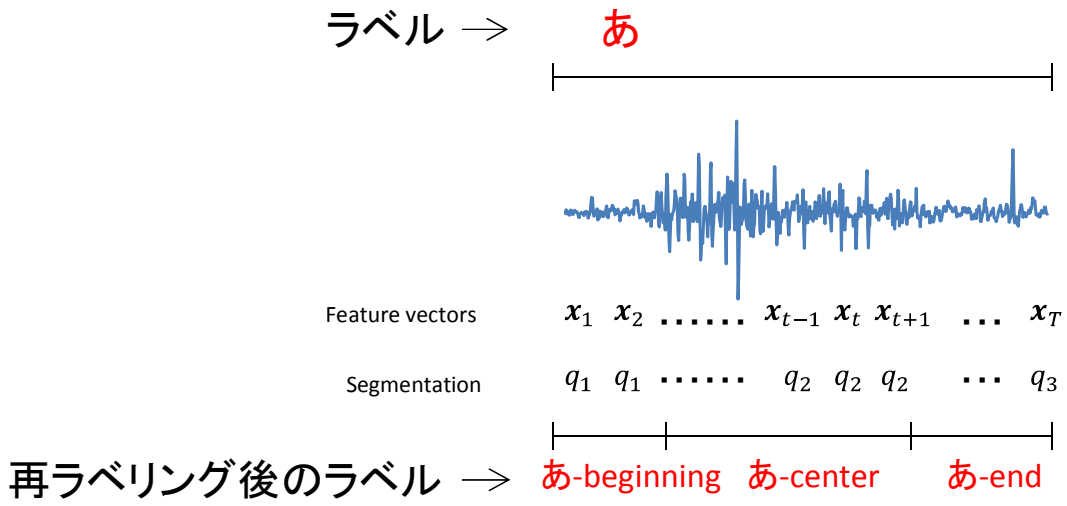

図 4.11: 「あ」の再ラベリング

第5章　日本語黙声認識実験

5.1　実験概要

本研究では，図5.1に示す日本語五十音のうち清音及び撥音の46音について黙声認識実験を行う．濁音に関しては，有声音の子音を含む音であるため，黙声認識においては無声音との認識はできないものと考えられる．

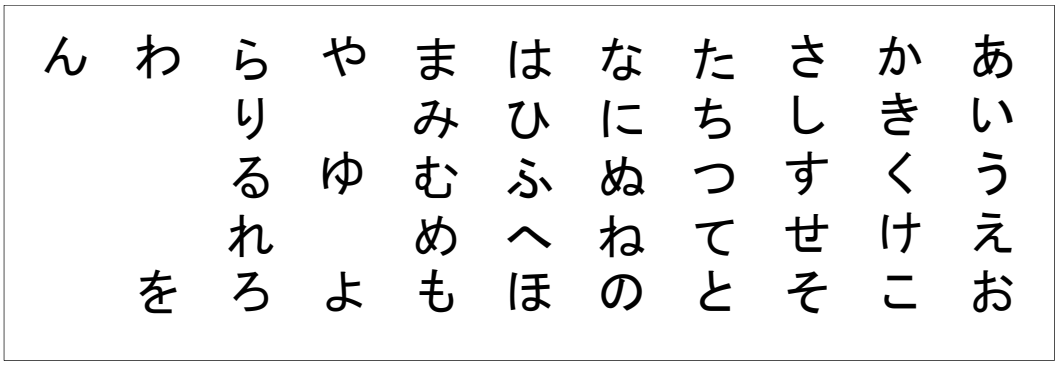

図 5.1: 日本語五十音のうち清音及び撥音の46音

そして，黙声認識におけるHybrid HMM/SVMを利用することの有効性を検証するため，二つの実験を行った．

実験1：日本語五母音に基づく認識

46音について，以下のようにそれぞれの音が含む母音を正解として認識する実験を行った．例えば,「か」は「あ」,「め」は「え」,「ほ」は「お」を正解とするラベルをつけるというように，各段ごとでの認識とした．撥音については認識対象音に含めなかった．

表 5.1: 実験1で用いた音の一覧とその正解ラベル

正解ラベル	認識対象の各段の音
あ	あ　か　さ　た　な　は　ま　や　ら　わ
い	い　き　し　ち　に　ひ　み　　り
う	う　く　す　つ　ぬ　ふ　む　ゆ　る
え	え　け　せ　て　ね　へ　め　　れ
お	お　こ　そ　と　の　ほ　も　よ　ろ

日本語五十音では，先行研究 [12][13] で確認されているように，主に母音に係る筋電位が観察され，子音に関する筋電位は母音の前に存在する．しかし，子音と母音は連続して発話されるため，音声における調音結合と同様に，観測される信号では分離できないことが考えられる．黙声認識で母音の認識が行われ，子音の検討があまり行われていない主な理由がこれに相当すると考えられる．そこで，先行研究 [7] を参考にして，上記の母音 5 クラスに関して認識を行う実験を行うことで，Hybrid HMM/SVM が時系列パターンの認識に有効であるか確認することを目的とした．

実験 2：日本語五十音に基づく認識

46 音をそれぞれの音として 46 音に認識する実験を行い，子音を含む日本語五十音の黙声認識における Hybrid HMM/SVM の評価を行った．この実験 2 では実験 1 の各段ごとの認識と異なり，子音の特徴を捉えることができなければ認識できない．そのため，認識可能な子音のパターンを発見するとともに，子音を表す短い時間パターンを Hybrid HMM/SVM で表現できるかがポイントとなった．

5.2　実験システム

Hybrid HMM/SVM による日本語黙声認識システムを図 5.2 に示す．実験 1 と実験 2 では，システムの構成自体に違いはなく，学習の際のラベリングにおける正解ラベルのみが異なる．

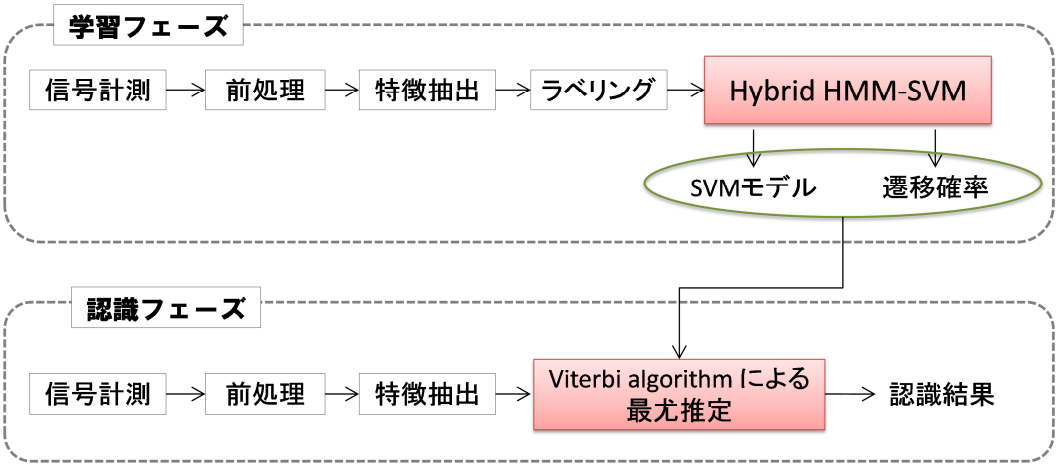

図 5.2: システム全体の構成

5.2.1　信号計測と前処理

被験者は 1 名で，筋電位の計測には湿式表面電極を用いた．電極の装着位置を図 5.3 に示す．計測位置に関しては，従来研究を参考にして，特に調音に深く関わる舌の動きに関与する筋に注目して選択した．Ch.1 の顎二腹筋は舌骨の引き上げ及び下顎の引き下げを行

う．Ch.2 の胸骨舌骨筋は舌骨の制御と開口運動に関わる．Ch.3 の広頸筋は口角の引き下げを行う．Ch.4 の口輪筋は口唇の突出に関与する．Ch.5 の大頬骨筋は口角を上方かつ外側に釣り上げる．参照電極は乳様突起（耳の裏側の付け根にある突起）上に装着した．

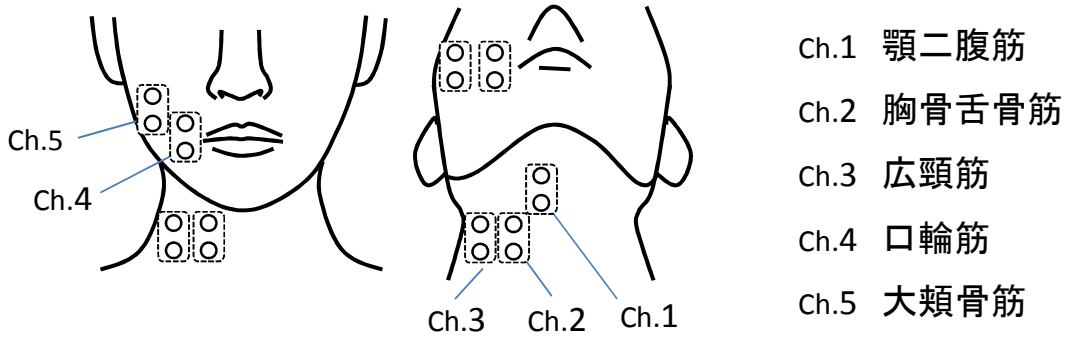

図 5.3: 筋電位の計測位置

被験者は電極を装着して椅子に座り，1 回の試行において約 1 秒間に 1 音の黙声発話を行うことを 46 音について繰り返し行った．1 音の発話では，筋を弛緩した状態から黙声発話を行い，再び弛緩状態に戻るまでを一つの発話区間とした．そして，10 試行分のデータを実験で用いた．

各チャネルで検出した筋電位信号に対して，筋電位計測装置内臓のハイパスフィルタ（遮断周波数 10[Hz]）をかけ，1000 倍に増幅して EMG を計測した．さらに，EMG を全波整流し，ローパスフィルタ（遮断周波数 2.4[Hz]）で平滑して IEMG とした．EMG，IEMG は，A/D 変換器を用いて，サンプリング周波数 1[kHz]，16[bit] でサンプリングし，計算機に取り込んだ．使用したハードウェア構成の詳細を表 5.2 に示す．

表 5.2: ハードウェアの構成

表面筋電位計測装置	Personal-EMG,（有）追坂電子機器
A/D 変換器	NI USB-6218，National Instruments（株）
計算機	CPU: Core i7-2700K 3.50[GHz]
	メモリ: 8.00[GB]
	OS: Windows 7

5.2.2　特徴抽出

第 3 章で述べたフレーム内平均 IEMG，ケプストラム係数を用い，ケプストラム係数は低次 3 項までとした．フレーム化に関しては，分析フレーム幅を 64[ms]，フレームシフト幅を 16[ms] とした．各チャネルに対して 1 次元のフレーム内平均 IEMG，3 次元のケプストラム係数を抽出して，各フレームで 20 次元の特徴ベクトルを抽出した．

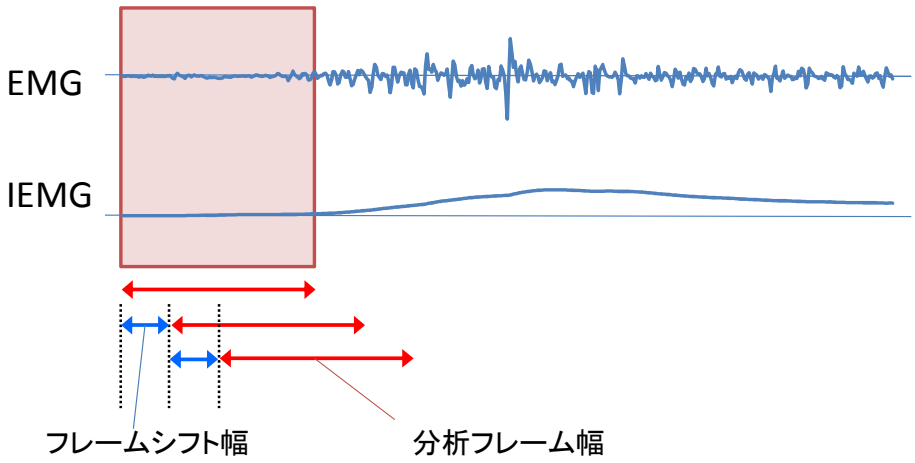

図 5.4: フレーム化

5.2.3　学習と認識

Hybrid HMM/SVM により，認識対象とする正解ラベル分の音をそれぞれモデル化する．実験 1 では日本語母音の 5 クラス，実験 2 では 46 音の 46 クラス分のモデルを作成することになる．SVM の連鎖構造は 3 状態 Left-to-Right HMM と同様にした．Viterbi アラインメントのための HMM についても 3 状態 Left-to-Right HMM とし，各状態の出力確率の密度関数は正規分布（対角行列，混合数 1）で学習を行った．そして，Viterbi アラインメントによって，各音について立ち上がり状態（beginning），定常区間（center），立下り状態（end）の 3 状態を得ることができたら，これを用いて SVM の学習を行った．SVM は RBF カーネルを用いた非線形 SVM とした．SVM のパラメータ C と RBF のパラメータ γ は，探索範囲 $C = \{2^0, 2^2, ..., 2^8\}$，$\gamma = \{2^{-5}, 2^{-4}, ..., 2^2\}$ として，学習データを用いて交差検定により最適値を求めた．

Hybrid HMM/SVM の他に，従来研究で用いられていた HMM と SVM を比較のために利用した．HMM に関しては，3 状態 Left-to-Right HMM で各状態の出力確率の密度関数は正規分布（対角行列，混合数 1）として，これも Hybrid HMM/SVM における構造と同様のものを用いた．また，SVM は，Hybrid HMM/SVM における SVM の構成と同様とした．ただし SVM では，1 フレーム単位で逐次認識処理が行われるため，入力とする一つの発話区間に対して一つの認識が得られるように投票処理を行って決定した．

5.2.4　評価方法

全 10 試行のうち，9 試行を学習データ，残りの 1 試行をテストデータとしてテストを行い，テストデータを変えてすべてのデータをテストする 10 重交差検定により評価を行った．実験 1 では各クラスの学習データの数に差があるため，データ数の少ない「い」段と「え」段のクラスに合わせるようにデータを間引いて学習を行った．

評価の指標として，音別認識率と平均認識率を用いる．音別認識率は，次の式により算

出する．

$$\text{音別認識率 [\%]} = \frac{\text{正解対象音数}}{\text{全試行内の対象音数}} \times 100 \tag{5.1}$$

平均認識率は，全認識対象音の平均値とした．

5.3　実験 1：日本語五母音に基づく認識の結果

実験 1 の結果として平均認識率を図 5.5 に示す．

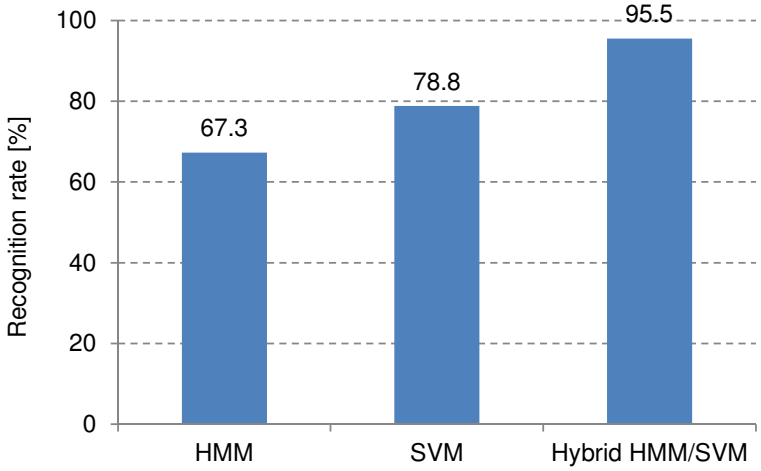

図 5.5: 実験 1 の平均認識率

Hybrid HMM/SVM による認識では，従来の HMM，SVM による認識に比べて高い性能が得られている．また，HMM，SVM，Hybrid HMM/SVM それぞれの場合の音別認識率を以下に示す．Hybrid HMM/SVM では，音別認識率がすべて 90% を超えるのに対して，HMM，SVM では誤認識が多くみられる．さらに,「か」をテストデータとした場合の,「あ」の Hybrid HMM/SVM モデルに対する Viterbi アラインメントを図 5.6 に示す．筋の弛緩状態から発話の立ち上がり，定常区間，立下りの 3 状態にセグメンテーションが行われていることが分かる．これにより，Hybrid HMM/SVM による認識が，黙声認識のような時系列データを扱う問題について有効であることが確認できたといえる．

表 5.3: 実験 1 の音別認識率（HMM）

		Recogniton results					
		あ	い	う	え	お	認識率 [%]
Spoken words	あ	80					100.0
	い		47	8	14	11	58.8
	う		4	58	11	7	72.5
	え		19	14	29	18	36.3
	お		5	13	7	55	68.8

表 5.4: 実験 1 の音別認識率（SVM）

Spoken words	Recogniton results: あ	い	う	え	お	認識率 [%]
あ	80					100.0
い		78	2			97.5
う		14	65		1	81.3
え		33		47		58.8
お		28	7		45	56.3

表 5.5: 実験 1 の音別認識率（Hybrid HMM/SVM）

Spoken words	Recogniton results: あ	い	う	え	お	認識率 [%]
あ	77	3				96.3
い	4	72	4			90.0
う		1	79			98.8
え				75	5	93.8
お				1	79	98.8

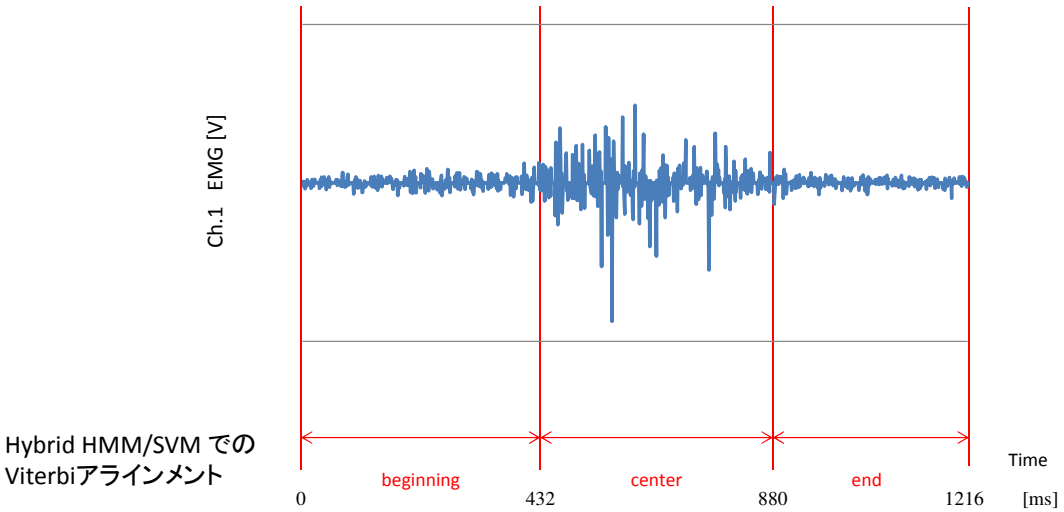

図 5.6: テストデータ「か」に対する Hybrid HMM/SVM「あ」での Viterbi アラインメント

5.4　実験 2：日本語五十音に基づく認識の結果

実験 2 の結果として平均認識率を図 5.7 に示す.

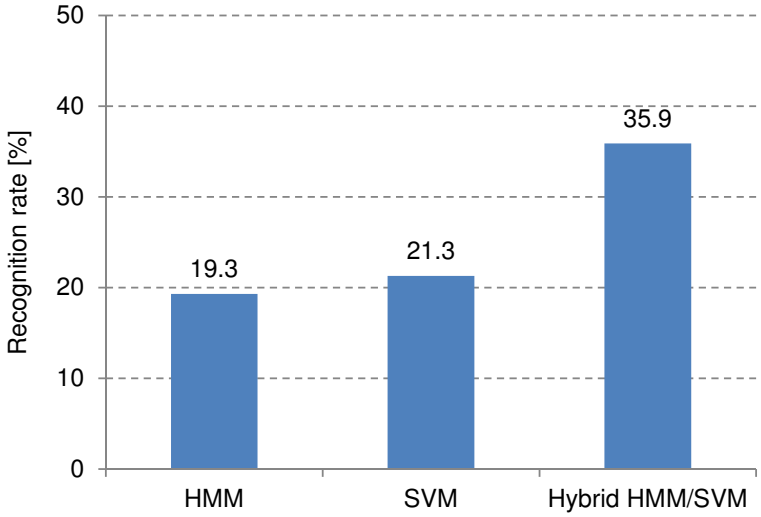

図 5.7: 実験 2 の平均認識率

Hybrid HMM/SVM による 46 音の黙声認識では，従来の HMM，SVM による認識に比べて高い認識率が得られている．この点で，Hybrid HMM/SVM が，従来手法である HMM，SVM と比較して黙声認識において有効であるといえる．しかし，平均認識率は 35.9% であり，黙声認識のためのシステムとしての性能は十分に高いとはいえない．図 5.8 に Hybrid HMM/SVM による音別認識率を示す．これをみると，各音に含まれる母音が同じものの間で誤認識が多く発生していることが分かる．これは，実験 1 でも見たように，各音の筋電位が主に母音に係るものであることが起因すると考えられる．各音に含まれる子音別の認識率を図中の最右列に示したが，子音の調音位置が近く，舌尖や舌端を用いる「さ」「た」「な」に関しては，認識率が 20% 以下となって他に比べて低くなっている．このことから，これらの音に関しては本実験での筋電位の計測位置では，類似したパターンしか得られていないことが考えられる．つまり，本実験では利用していなかった筋の計測を検討することで，これらの認識も可能となる可能性はあるが，表面筋電位では浅い部分にある筋に関する信号しか基本的には観測することができないため，これ以上の情報を取得することは難しいといえる．

そこで，今後の展開としては，調音の方法が似ている子音をグループ化して認識を行うことで，認識可能な子音を探し出すことが考えられる．また通常の音声認識では，連続して発声した場合の前後の音の組み合わせを考慮して一つのモデルとし，各音素のトライフォンモデルを作成する方法で認識精度を向上させることが行われており，Hybrid HMM/SVM でも同じ手法でモデルを作成することによって，認識精度を上げることも考えられる．さらに，本実験では被験者の人数や試行数が十分に収集することができなかったため，非常に少ないデータの中でも Hybrid HMM/SVM による認識が，従来の HMM や SVM よりも高い性能を発揮することが確認できたものの，より多くのデータを集めて Hybrid HMM/SVM の効果についてもさらに検証する必要がある．

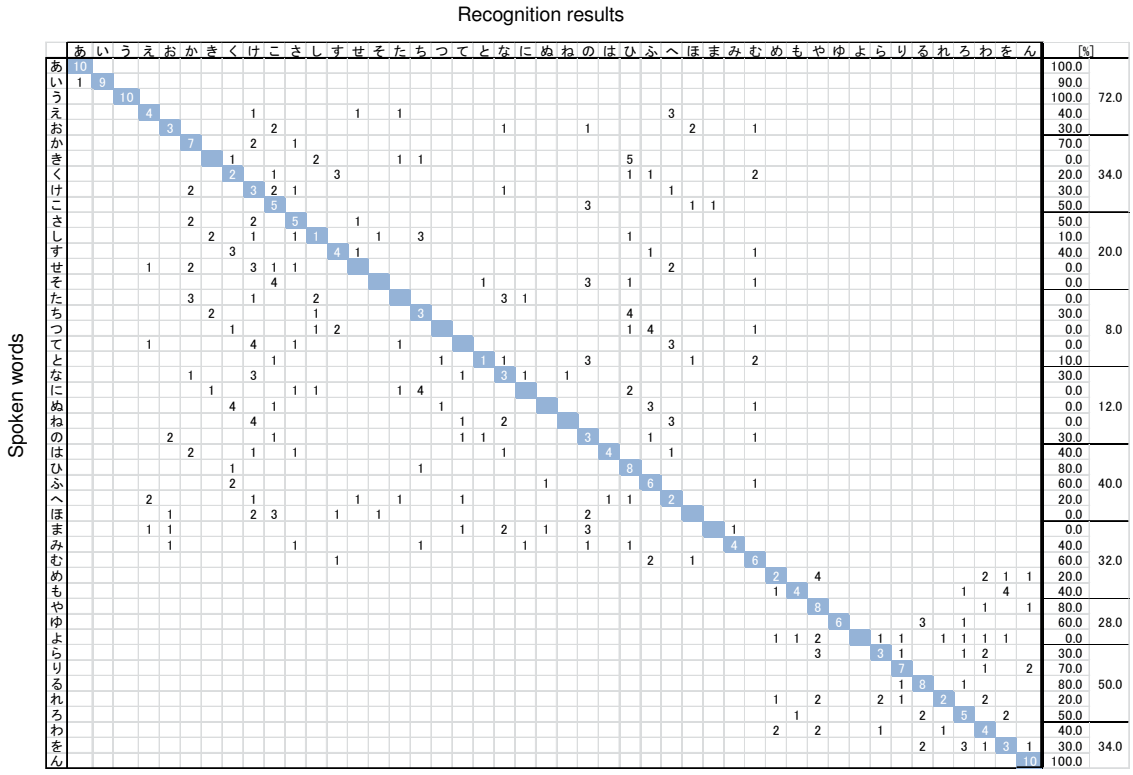

図 5.8: 実験 2 の音別認識率（Hybrid HMM/SVM）

第6章　結論

本研究では，日本語で用いられる母音と子音を組み合わせた五十音を認識することを目的として，従来の認識手法を改良した Hybrid HMM/SVM を利用することを提案し，評価実験による検討を行った．本研究で提案する Hybrid HMM/SVM では，汎化性能に優れた SVM を HMM のように連鎖させて利用し，SVM を時系列で伸縮するパターンに対して適応することで，データを大量に用意することができない場合でも認識が可能である．

日本語五十音の黙声認識を行う評価実験では，表情筋及び頸部の 5 箇所で筋電位を計測し，五十音のうち清音と撥音の 46 音のデータを用いて認識実験を行った．実験では，まず 46 音について，それぞれの音が含む母音を正解として認識させた．例えば,「い」段の「き」は「い」,「し」は「い」，また,「え」段である「め」は「え」などとした．このとき，Hybrid HMM/SVM による平均認識率は 95.5% で，従来手法である HMM での 67.3%，及び SVM での 78.8% と比較して優れた認識性能が得られた．次に，46 音をそれぞれの音（例えば,「き」は「き」,「し」は「し」,「め」は「め」など）として認識する実験を行った結果は，Hybrid HMM/SVM での認識率が 35.9% となり，HMM での 19.3%，SVM での 21.3% に比べて高い認識率を得た．これらの実験から，黙声認識における Hybrid HMM/SVM の利用が，従来手法に比べて有効であることが確認できた．

46 音を各音として認識した結果を分析して,「あ」「か」「は」「ま」「や」「ら」行の音と，それ以外の行の音での認識率を比較すると，子音の調音位置が近く，舌尖や舌端を用いる「さ」「た」「な」に関しては，認識率が 20% 以下となって他に比べて低くなっている．このことから，音声学における子音の調音位置による分類が影響していることが示唆される．つまり，調音のための動作を示す筋電位を取得できていたものの，その動作自体が類似していたために認識率が悪い行があったと推測される．

今後は，子音をグループ化することでその認識可能な範囲を検討することが考えられる．また，本研究では計測データの数が先行研究に比べて少なかったため，より多くのデータを収集しながら認識モデルの効果をさらに検証する必要がある．

謝辞

学類 4 年次から御指導くださいました三河正彦准教授に心から感謝致します．研究テーマの変更や就職活動においても親身になって相談に乗ってくださいました．

また，研究活動に関して貴重な御指摘をいただいた田中和世教授，藤澤誠助教授に御礼申し上げます．

産業技術総合研究所の児島宏明氏，吉川雅博氏には研究に関する御助言を多々いただいたりとお世話になりました．特に，三河研究室の先輩で，本研究の先行研究を行っておられた吉川雅博氏には，お忙しい中多くのノウハウを授けていただきましたこと，深く感謝致します．

研究室で共に過ごした田中研究室，藤澤研究室，三河研究室の皆様，いつも支え，励ましてくれた中沢彰吾君，大学入学時から様々な面で助けてくれた石崎琢弥君，そして最後に，これまで支え，応援してくれた家族に心から感謝致します．

大内　慶久

2013 年 3 月

参考文献

[1] ”Voice Search - Google,” http://www.google.com/insidesearch/features/voicesearch/index.html, (accessed 2013-01-15).

[2] ”Use your voice to do even more with Siri - iOS - Apple,” http://www.apple.com/ios/siri/, (accessed 2013-01-15).

[3] 中田康之, 安藤護俊, ”色抽出法と固有空間法を用いた読唇処理”, 電子情報通信学会論文誌, Vol.J85-D-II, No.12, pp.1813-1822, 2002.

[4] 齊藤剛史, 石倉寛之, 山下晃平, 小西亮介, ”トラジェクトリ特徴量を利用した単語読唇に関する基礎検討”, 電子情報通信学会技術研究報告 HIP, Vol.109, No.471, pp.259-264, 2010.

[5] 真鍋宏幸, 平岩明, 杉村利明, ”筋電信号を用いた無発声音声認識 —定常状態における母音の識別—”, 情報処理学会シンポジウム論文集, Vol.2002, No.7, pp.181-182, 2002.

[6] 真鍋宏幸, 平岩明, 杉村利明, ”無発声音声認識:筋電信号を用いた声を伴わない日本語5母音の認識”, 電子情報通信学会論文誌, Vol.J88-D-II, No.9, pp.1909-1917, 2005.

[7] 福田修, 藤田真治, 辻敏夫, ”EMG 信号を利用した代用発声システム”, 電子情報通信学会論文誌, Vol.J88-D-II, No.1, pp.105-112, 2005.

[8] 張志鵬, 真鍋宏幸, 堀越力, 杉村利明, ”HMM 及びケプストラム係数特徴による筋電信号を用いた無発声音声認識”, 電子情報通信学会技術研究報告, Vol.103, No.401, pp.7-12, 2003.

[9] 吉川雅博, 児島宏明, 田中和世, ”頸部から計測した筋電位信号を利用した発話認識”, ヒューマンインタフェース学会論文誌, Vol.11, No.3, pp.293-302, 2009.

[10] C. Jorgensen and K. Binsted, ”Web Browser Control Using EMG Based Sub Vocal Speech Recognition,” in Proc. of the 38th Hawaii International Conference on System Sciences, 2005.

[11] Luay Fraiwan, Khaldon Lweesy, Ayat Al-Nemrawi, Sondos Addabass, Rasha Saifan, ”Voiceless Arabic Vowels Recognition using Facial EMG,” Med. Biol. Engineering and Computing, Vol.49, No.7, pp.811-818, 2011.

[12] 菅沼健, 森博彦, ”筋電位信号による子音認識のための認識パラメータの調査”, 情報処理学会第 73 回全国大会講演論文集, pp.115-117, 2011.

[13] 永井秀利, 南誠子, 中村貞吾, 野村浩郷, ”筋電に基づく黙声認識における子音認識のための基礎的調査”, 電気関係学会九州支部連合大会 12-1P-06, 2004.

[14] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel, ”Session Independent Non-audible Speech Recognition Using Surface Electromyography,” Automatic Speech Recognition and Understanding, Puerto Rico, 2005.

[15] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel ”Towards Continuous Speech Recognition Using Surface Electromyography,” in Proc. ICSLP 06, Pittsburgh, USA, 2006.

[16] Hiroaki Kawamoto, Yoshiyuki Sankai, ”EMG-based Hybrid AssistiveLeg for Walking Aid using Feedforward Controller,” Proc. of ICCAS2001IEEE International Conference on Control, Automation and system, pp.193-194, 2001.

[17] 中井浩一, 中井満, 下平博, 嵯峨山茂樹, ”Support Vector Machine による時系列パターンの認識”, 電子情報通信学会技術研究報告 PRMU, Vol.99, No.514, pp.15-20, 1999.

[18] Jaume Padrell-Sendra, Dario Martin-Iglesias, Fernando Diaz-de-Maria, ”Support vector machines for continuous speech recognition,” 14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, 2006.

[19] A. Ganapathiraju, J. Hamaker, J. Picone, ”Hybrid SVM/HMM Architectures for Speech Recognition,” in Proc. of the International Conference on Spoken Language Processing, vol.4, pp.504-507, 2000.

[20] 吉川雅博 , 三河正彦 , 田中和世, ”筋電位を利用したサポートベクターマシンによる手のリアルタイム動作識別”, 電子情報通信学会論文誌, Vol.J92-D, No.1, pp.93-103, 2009.

[21] 井部鮎子, ”手の動作識別を目的とした筋電位信号からの特徴抽出手法”, 筑波大学図書館情報専門学群卒業論文, 2007.

[22] 奈良勲, 岡西哲夫, ”筋力”, 医歯出版株式会社, 2004.

[23] 木塚朝博, 増田正, 木竜徹, 佐渡山亜兵, ”表面筋電図”, 東京電機大学出版局, 2006.

[24] Peter Ladefoged, 竹林滋（訳）, 牧野武彦（訳）, ”音声学概説”, 大修館書店, 1999.

[25] ”SVM Application List,” `http://www.clopinet.com/isabelle/Projects/SVM/applist.html`, (accessed 2013-01-15).

[26] 荒木雅弘, ”フリーソフトでつくる音声認識システム”, 森北出版, 2007.

[27] ”HTK Speech Recognition Toolkit,” `http://htk.eng.cam.ac.uk/`, (accessed 2013-01-15).

[28] ”大語彙連続音声認識エンジン Julius”, `http://julius.sourceforge.jp/`, (accessed 2013-01-15).

[29] Donald Knuth, ”Two Notes on Notation”, American Mathematical Monthly, Volume 99, No.5, pp.403-422, 1992.

[30] L. E. Baum, T. Petrie, G. Soules, N. Weiss, ”A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains,” The Annals of Mathematical Statistics, Vol.41, No.1, pp.164-171, 1970.

[31] N. Cristianini, J. Shawe-Taylor, 大北剛（訳）, ”サポートベクターマシン入門”, 共立出版, 2005.

[32] J. Platt, ”Fast Training of Support Vector Machines Using Sequential Minimal Optimization,” in Advances in Kernel Methods: Support Vector Learning, B. Schölkopf, C. Burges, A. Smola, eds., pp.185-208, MIT Press, 1999.

[33] J. Platt, ”Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods,” in Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans, eds., MIT Press, 1999.

[34] 河合良訓, 原島広至, ”肉単”, NTS, 2004.

[35] 長谷川智紀, ”複数マイクロホンを用いた雑音低減法の音声認識への応用”, 筑波大学大学院図書館情報メディア研究科修士論文, 2004.

[36] 中村匡伸, ”話し言葉音声の音響的・言語的特徴に関する研究”, 東京工業大学大学院情報理工学研究科博士論文, 2009.

[37] 井口茂, ”類似動作からの特徴部位抽出 - 人間の行動認識を例として -”, 早稲田大学理工学部情報学科卒業論文, 2005.

[38] M. Wand, T. Schultz, ”Analysis of phone confusion in EMG-based speech recognition”, in Proc. ICASSP, pp.757-760, 2011.